DOCUMENT RESUME

ED 068 518 TM 001 879

AUTHOR	Jennings, William N.
TITLE	The Reliability and Validity of the Word-Association Test as a Measure of Academic Achievement. Final Report.
INSTITUTION	National Center for Educational Research and Development (DHEW/OE), Washington, D.C.
SPONS AGENCY	Office of Education (DHEW), Washington, D.C.
BUREAU NO	Proj-1-C-043
PUB DATE	Mar 72
GRANT	OEG-3-71-0115
NOTE	37p.
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	*Academic Achievement; *Associative Learning; Cognitive Processes; Data Processing; *Educational Research; *Measurement Techniques; Scoring; Stimulus Devices; Testing; Test Reliability; Tests; Test Scoring Machines; *Test Validity; Word Lists

ABSTRACT

A simply constructed, psychometrically sound testing procedure which enables the instructor to assess higher cognitive process with respect to the material in questions, and which is amenable to machine scoring, is described. It involves the application of the word association technique long used in psychoanalysis to the classroom setting. The actual testing procedure requires the selection by the instructor of a number of stimulus terms which sample a wide range of the concepts covered in the course. The student is usually required to produce four associatives for each stimulus term, providing sufficient discriminatory power for evaluating student knowledge. After details of the method, its analysis and psychometrics are given, it is concluded that the word-association technique is of sufficient reliability and validity to warrant further investigation. Nearly the entire second half of the report is devoted to appendices on instructor orientation material, lists of stimulus items for each course, sample test, word count, psychometric characteristics, and reliability calculation. (LH)

ED 068518

APR 1972
OE - NCERD
TM
PROJ 1-C-043

Final Report

Project No. 1-C-043
Grant No. OEG-3-71-0115

William N. Jennings
Randolph-Macon College
Ashland, Virginia 23005

THE RELIABILITY AND VALIDITY OF THE WORD-ASSOCIATION TEST AS A MEASURE OF ACADEMIC ACHIEVEMENT

March, 1972

U.S. DEPARTMENT OF HEALTH EDUCATION AND WELFARE

Office of Education

National Center for Educational Research and Development
(Regional Research Program)

TM 001 879

ED 068518

Final Report

Project No. 1-C-043
Grant No. OEG-3-71-0115

THE RELIABILITY AND VALIDITY OF THE WORD-ASSOCIATION TEST AS A MEASURE OF ACADEMIC ACHIEVEMENT

William N. Jennings

Randolph-Macon College

Ashland, Virginia

March 29, 1972

The research reported herein was performed pursuant to a grant with the Office of Education, U.S. Department of Health, Education, and Welfare. Contractors undertaking such projects under Government sponsorship are encouraged to express freely their professional judgment in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent official Office of Education position or policy.

U.S. DEPARTMENT OF
HEALTH, EDUCATION AND WELFARE

Office of Education
National Center for Educational Research and Development

TABLE OF CONTENTS

I. Introduction..... 1

II. Problem..... 5

III. Method..... 6

IV. Analysis..... 7

V. Psychometrics..... 8

VI. Results..... 9

VII. Discussion..... 10

VIII. Conclusions..... 14

IX. References..... 16

X. Appendixes..... 18

Appendix A..... 18

Appendix B..... 19

Appendix C..... 24

Appendix D..... 27

Appendix E..... 29

Appendix F..... 31

LIST OF TABLES

TABLE		PAGE
I.	Summary of Disciplines, level , Number of Students.....	6
II.	Grand Means Item Difficulty.....	9
III.	Reliability and Validity Estimates....	10
IV.	Stimulus Items, Mean and Std. Deviations Item Difficulty.....	19
V.	Word Association Psychometric Characteristics.....	29
VI.	Word Association Test Validity Coefficients.....	30

LIST OF FIGURES

FIGURE		PAGE
I.	Association Test Format.....	3
II.	Association Test Scoring.....	4
III.	Word Count Format.....	4
IV.	Instructors Scoring List.....	5

Introduction:

A cursory review of the literature on educational measurement and evaluation will reveal that few if any really new techniques for measuring academic achievement have been developed in the last twenty-five years. The time honored testing techniques, multiple choice, essay, matching, etc, have been widely used for some time and are described in virtually all the measurement texts old or new. (Remmers, Gage, and Rummel, 1965). Certainly methods for improving selection and/or development of testing materials have evolved (Mager, 1962; Payne, 1968), but the test formats themselves remain rather static. The long standing debate between proponents of objective and subjective methods has failed to produce any new methods (Noll, 1965). Current enrollment trends (Carter, 1967; Milton, 1968) indicate larger and larger classes with a necessary concommitant shift to objective examinations to facilitate scoring. To be sure, objective examinations may be developed which sample such high level cognitive processes as synthesis, integration, and evaluation (Bloom, 1956). It is also true that the majority of teacher-made items bear little resemblance to those proposed by Bloom, and that the time required to develop such items is often prohibitive.

What is needed then is a simply constructed, psychometrically sound, testing procedure which enables the instructor to asses higher cognitive process with respect to the material in questions, and which is amenable to machine scoring.* The testing procedure to be described herein appears, after fairly extensive preliminary research, to be just such a device.

This promising new procedure, conceived by Dr. W.S. Verplanck at the University of Tennessee and undergoing continuous developmental research there, involves the application of the word-association technique long used in psychoanalysis (Woodworth and Schloshberg, 1960) to the class-room setting. While the testing procedure bears superficial resemblance to the established clinical technique it is not from this source, but from contemporary research in concept formation, memory, psycholinguistics, and human thought processes that the academic measurement application was derived.

---

*While machine scoring was not included in this project, preliminary steps toward ultimate computer scoring of the word association exam have been taken at the University of Tennessee.

That word-associations play a central role in current psychological theorizing in these areas can be easily shown. Creelman (1966) in her review of research 'meaning' concludes that, "surely associations between and among words must play a large and important role in any adequate definition of meaning." Deese (1965) sums up his position on the study of associations as follows:

> "We study associations in order to make inferences about the nature of human thought, and these associations are cast in the language which embodies the thought . . . To the extent that verbal behavior is the mediator of thought, modern association theory is the theory of thought. The whole of current concern with associatives mediators, as a matter of fact, is an effort to use the associative properties of explicit verbal behavior as a model for the implicit verbal thought processes."

In a more applied sense Underwood and Richardson (1956) and Freedman and Mednick (1958) demonstrated that verbal concept attainment is a function of the underlying associative responses involved. Verplanck (1962) has demonstrated that the hypotheses generated by a concept formation subject are a function of the available associative links between stimulus items. Bousfield (1953) and many others have shown that recall of word lists is facilitated by the presence of shared association.

The actual testing procedure requires the selection by the instructor of a number of stimulus terms which sample a wide range of the concepts covered in the course. Tests are constructed using these items in the format shown in Figure 1. The vertical array using small boxes was devised when early horizontal formats were found to generate sentences rather than the preferred word or phrase, perhaps due to their resemblance to normal left to right cursive writing, and also because the horizontal format lent itself to response chaining (Verplanck, 1968). The typical test consists of ten four response items per 8 1/2 by 11 page. The student is usually required to produce four associatives for each stimulus term, although this number can vary as circumstances require. Early research indicated that four associatives would in most cases provide sufficient discriminatory power for evaluating student knowledge. Unless the concept has been covered in considerable detail requiring more than four responses leads to diminishing returns. In general a four response item will require approximately one minute to answer. An hour test consisting of 45 to 50 items can usually be completed by the average student.

Binet	Genius	Twin Study
Binet	Genius	Twin Study
Binet	Genius	Twin Study
Binet	Genius	Twin Study

Figure 1. Association test format.

The instructions which accompany the test have varied, particularly with respect to their emphasis on the production of a single word in response to the stimulus item. In general strong emphasis on a single word or phrase coupled with the vertical format seems to produce the best results. The instructions used in this study were as follows:

> "In the space alotted on the mimeoed page, briefly present those associations which you make to each item whcih are most directly relevant to the subject matter of (psychology). A "pinpointing" word, or phrase, is all that is necessary to demonstrate that you know what you're writing about, and that you could write a lot more, if given the time. Do not ever repeat a word or term in response to the same item. Don't Guess!

Scoring of the word-association test is generally based on the four-point scale shown in figure 2. The majority of the responses will fall in either the +2 or the 0 category. Negative scores are surprisingly rare. Using this four point scal the potential range of the test, assuming four responses per stimulus term, is from minus four times the number of items to plus eight times that number. A twenty item test would have a potential range of 240 points, from -80 to +160. The wide range of scores obtained from relatively few items gives the test excellent discriminatory power. A simple binary routine where each response is judged acceptable or unacceptable has also been employed, but this reduces the discriminatory power of the examination.

The word count format shown in figure 3 plays an important role in the scoring process. Responses are counted clerically, and a complete list of all responses to each stimulus word is compiled. From this list an alphabetized list of all unique responses (see figure 4, page 5) is

presented to the instructor for scoring. A response that occurs 20 times is, thus scored only once. Within grader reliability is perfect and objectivity is insured. Once the instructor has assigned values to the associates on his scoring sheet, the list becomes a dictionary for the clerical scoring of the individual word-association examinations.

Stimulus Word: Binet

Response	Score	Rationale
I.Q. Test	+2	A good association, relevant, demonstrates grasp of material.
French Psychologist	+1	Acceptable but not informative a reasonable guess.
Stephen Vincent	0	Out of context, irrelevant
Freuds' Student	-1	Positively incorrect

Figure 2. Association test scoring

Source:
Date:
Stimulus Word: Binet

Associations	Score	Position				Total Freq.	Total Score
		1	2	3	4		
Henri	2	1111	1111	111	1	13	26
Simon	2	1	1			2	4
I.Q. Test	2	11	1			3	6
French	1				1	1	1
Steven Vincent	0			1		1	0
Wundts' Student	-1				1	1	-1

Note: Score and total score columns are added after scoring is completed. Position refers to the ordinal position of the associate, (e.g. whether it was given first or fourth).
Figure 3. Word Count Format

Stimulus Word: Binet

Responses:	Score
French	1
Henri	2
I.Q. Test	2
Simon	2
Steven Vincent	0
Wundts' Student	-1

Figure 4. Instructors Scoring List

Research conducted prior to this project indicated that the word-association testing procedure is an extremely reliable one. Coefficients ranged from .724 (N = 35) to .943 (N = 28), the most stable estimates being .844 (N = 133) and .885 (N = 364). Validity estimates range from .733 (N = 25) to -.13 (N = 25) depending on the criterion. Most coefficients are between .45 and .65 and are thus well within the acceptable range.* Two problems exist with respect to this early research. Virtually all the data came from large sections of courses in psychology. It seems likely that the reliability is characteristic of the technique per se, and not the area tested, but there is little empirical support for this assertion. Since large sections were used, the validity estimates are almost entirely based on multiple choice criterion or the GPA, in this case a multiple choice derivative.

Intuitively, writing an essay requires that the individual produce the key terms (associates) and then string them together grammatically. The content bearing part of the essay would thus appear to be closely related to associative processes. If the close linkage exists between associations and essay content, then the correlations between word-association tests and essays tests over the same material should be uniformly high.

Problem:

The research reported herein, by administering the word-association test in nine different subject matter areas ranging from political science to biology, sought to verify empirically the generality of the reliability of the word-association technique. By administering these tests in conjunction with regular final examinations, which are predominantly essay at Randolph-Macon, the validity of the test with essay criteria was also investigated.

The specific hypotheses which this research sought to confirm are as follows:

---

*A complete summary is provided in Appendix E, page 30.

1. The word-association testing technique is highly reliable in a broad range of non-quantitative curricular areas.
2. The word-association test is related to the essay test in terms of the cognitive function measured. Correlations between the two will be positive and high.
3. The word-association technique is of sufficient reliability and validity to have great potential for educational measurement and research.

Methods:

Sample: Participating instructors were selected to include the widest range of non-quantitative disciplines in the testing program. Courses in which the tests were administered were selected to maximize the number enrolled, and, where possible, to represent the full range of course levels. As a result of this process some students were tested in more than one course. Thus the coefficients obtained are not entirely independent of each other. Table 1 provides a summary of disciplines, number enrolled and level of the courses included in this research.

Table 1

Summary of Disciplines, Level, Student Number

Course	Level	N
Psychology	Lower	29
Political Science	Upper	17
History	Upper	17
English	Lower	17
Philosophy	Lower	18
Economics	Lower	18
Education	Upper	32
Religion	Upper	24
Botany	Lower	56
Sigma		228

Instructor Orientation: Participating instructors were given a brief orientation to the association testing procedure. Item selection, test format, instructoions, scoring and timing were covered in detail, and instructors were encouraged to raise any questions which occurred to them. Instructors were provided with a written summary of the materials covered in orientation (see Appendix A).

Item Selection: Each instructor was asked to submit to the investigator a list of twenty terms which sampled key concepts covered in their course. Items were to cover the entire semester. The investigator met individually with the participants to assist in selection of items. Lists of all stimulus items for each course may be found in appendix B.

Test Construction and Format: Tests were constructed using the items submitted. Each test consisted of three pages; a cover page with instructions and an example of correct associative responding, and two ten item pages of stimulus terms arranged in the format shown on page 2. A sample test may be seen in appendix C, pages 24 . Instructions were identical for all participants.

Test Administration: Tests were prepared by the investigator and returned to the participating instructors in advance of their scheduled final examination dates. Each instructor administered the word-association test in his own course in conjunction with his final examination. In order to standardize the administration as much as possible instructors were asked to allow the first thirty minutes of the examination period for the association test. Since association items usually take approximately one minute each to answer, this provided ample time for the knowledgeable student. In order to maintain motivation instructors were asked to respond to the question as to whether the test 'counted' with "the test will be scored and the results returned to me." Completed examinations were turned over to the investigator. Scores on the regular final examinations were turned over to the investigator following their use in determining course grades.

Analysis

Word Count: Word-association tests were analyzed as follows. A word-count (format shown on page 4) was compiled clerically for every stimulus term on each test. From the word counts an alphabetical list of all unique responses was derived. Derivation of this list was facilitated by the development of a computer program which made accessible on a remote teletype an alphabetic sorting routine. A complete word-count and a scored instructors' list may be found in appendix D.

Scoring: Alphabetized lists of associatives were returned to the participating instructors for scoring on the +2 to -1 basis shown on page 5 . Each instructor scored the responses without knowledge of the individual who made the response, or of the context in which it occurred (since each response was scored only once, grader reliability was perfect). Lists scored by the instructor

were returned to the investigator. Using these lists as a dictionary scores were transposed to the response on the individual tests. Scores for each item, and a total score on the test were calculated for each participating student. These scores were coded on a loose leaf data sheet designed to facilitate key punching. Also included on these sheets were scores on the regular final examination, broken down into objective and subjective parts scores where appropriate, overall grade-point averages as of the preceeding semester, and an identification number. These data were punched onto IBM cards for the reliability and validity analysis.

Psychometrics

Item Difficulty: Item difficulty may be estimated on the word-association test by calculating the mean score for each stimulus term. The potential range of scores is from minus four to plus eight, but the effective range seems to be bounded by zero. Negative scores are relatively uncommon, while blanks are numerous. A grand mean of item difficulty can also be calculated as an estimate of the overall difficulty of the test. Lacking the usual pass-fail criterion, a difficulty score of four is an indication of moderate difficulty, while approaching two and six are difficult and easy respectively.

Reliability: Reliability estimates were calculated using the Kuder-Richardson Formula 20 as modified by Dr. E.E. Cureton for use with associative data. See appendix F.

Validity: Validity estimates were calculated using the product moment correlation coefficient. Three measures of word-association validity were obtained. Association test scores were correlated with objective and/or subjective final examination part scores, and with overall GPA. Subjective part scores were primarily essay, but also included identification, fill-in-the-blanks, and short answer. Objective part scores were derived entirely from multiple-choice items.

Programming: All calculations were made using the college's IBM 1800 computer. A fortram II program consisting of a driver and subroutines for reliability and validity was written by the investigator with help from the computer center staff. The print-out included a student by item matrix of association scores, mean scores per item (item difficulty) and per student and total score per student, the K.R. - 20 reliability coefficient, and a matrix of correlations indicating validity. Each section was analyzed as the cards were punched and as computer time was available. A documented copy of the program is available on request.

Results:

Item difficulty: Table 2 presents the grand means of item difficulty for each course. This figure represents the overall difficulty of the word-association examination. A complete summary of individual item difficulties may be found in Table 4, appendix B.

Table 2
Grand Means of Item Difficulty

Course	N	$\bar{X}$ Difficulty	S Difficulty	% Blank
Psychology	29	5.93	.86	6
Political Science	17	4.82	1.28	23
History	17	4.26	1.58	9
English	17	2.03	1.37	51
Philosophy	18	3.30	1.56	25
Economics	18	3.29	1.22	12
Education	32	2.35	.80	23
Religion	24	4.03	.83	10
Botany	50	4.96	.73	13

English with a mean difficulty score of 2.03 was by far the most difficult examination. This contention draws support from the fact that 51% of the total responses were blank. At the other extreme the Psychology test with a mean difficulty score of 5.93 was decidedly easy. Again supported by the finding of only 6% blanks. Tests in other courses tend to cluster around 4.00 (indicating moderate average difficulty) with per cent blank ranging from 9% to 25%, the mean being 13%.

Reliability: Table 3 presents the reliability and validity coefficients obtained in this study.

As can be seen in column 2 of the table the reliability coefficients range from .53 (N = 29) to .89 (N = 18). Seven of the nine coefficients obtained are consistent with the findings of previous research (Appendix E) and several are extremely high for short teacher-made tests. The two low coefficients are associated the extremes of average item difficulty, English and Psychology.

Validity: Column 3, contains the validity estimates obtained using subjective criterion. These correlations are generally smaller than the coefficients obtained in

earlier research (Appendix E). While seven of eight are in the predicted direction, only three reach statistical significance. The lone negative value is from English with its high mean item difficulty.

Validity estimates using overall grade-point-average as criteria are found in column 4 of Table 3. The values obtained range from -.21 to .59. These coefficients differ widely from the findings of previous research. Three of the five positive correlations are statistically significant, while none of the four negative values approach significance. Two of the negative values stem, once again, from the item difficulty extremes.

Column 4 Table 3 presents the correlations obtained between the subjective final examination score and GPA, Here six of the eight are statistically significant at or beyond the .05 level, and the two remaining are from English and Psychology, the item-difficulty extremes.

Table 3

Reliability and Validity Estimates

Course	N	Kr-20	Validity Assn vs S	Assn vs GPA	Sys GPA
Psych	29	.53	.30	-.15	.01
Poly Sci	17	.85	.35	-.21	.54*
History	17	.75	.52*	.53*	.48*
English	17	.56	-.39	-.10	-.17
Philosophy	18	.89	.73**	.56*	.61**
Econ	18	.84		.59**	
Educ	32	.70	.31	.24	.56***
Religion	24	.79	.21	.33	.59**
Botany	56	.86	.28*	-.08	.38**

The Economics final was entirely objective. The correlation between the association test and the final was -.05 and between the final and GPA .46.

* .05
** .01
*** .001

Discussion:

The results obtained in this study are sufficiently mixed to make discussion difficult. With regard to item difficulty, each participating instructor selected his

own items. Since the items were to sample key concepts from the material covered by the course, it was not possible for the investigator to pass judgement on the items to which the informed student could provide four good associative responses, but the range of item difficulties obtained indicates that different interpretations of the term 'informed' prevailed. Since instructors also scored the responses, it is also possible that the items were appropriate, but that the scoring was either too conservative or too liberal. No 'hard and fast' solution to this problem suggests itself. Since the instructors knowledge of course content and emphasis is unique to him, he alone is in a position to judge the value of the responses.

Despite these limitations seven of the nine participants were able to select items which were of moderate difficulty on the average. Further, these instructors had no prior experience with the word-association testing procedure. This suggests that item difficulty extremes will rarely present problems for the individual using this technique, but the effect of such extremes on reliability and validity makes caution necessary.

As expected item difficulty exerts a strong influence on Reliability. The reliability coefficients obtained in this study are generally high. The two low coefficients can be traced to item difficulty averages which indicate that one test (English) was too difficult, and the other (Psychology) too easy. In English 51% of the total responses were blanks. If the majority of students simply fail to answer the questions the test cannot measure, let alone measure reliably. It is quite possible to develop a difficult test which measures reliably. In Education the mean item difficulty was 2.35 SD(.80) which is quite similar to English, but the percentage of blank spaces was .23%, less than half as many. In English the items were so difficult as to discourage responding, in Education the responses were scored conservatively.

At the other end of the continuum only 6% of the total responses were left blank in psychology. Here, almost everyone received full credit for every response listed. Under these circumstances the test fails to discriminate, and the reliability is reduced accordingly. Again, one can design an easy test in terms of the students willingness to answer which will measure reliably. In History only 9% of the responses were left blank, but the mean item difficulty was moderate (4.26). and the reliability was .75.

For those sections with moderate mean difficulties and a reasonable proportion of blanks the reliability

coefficients were at least satisfacotry, with several being quite high. These coefficients are on the same order as those previously obtained and seem to confirm the hypothese that the procedure is reliable in a broad range of subject matter areas.

Item difficulty also exerts an influence on validity, in essence, if a test fails to measure it can be neither reliable nor valid. Validity estimates suffer from a second problem too. The small number of students enrolled in each course makes it necessary to obtain a very high correlation (as validity coefficients go) to achieve statistical significance. This same small number makes feasible the use of essay examinations. No solution for this paradoxical situation suggests itself.

The results with respect to validity are difficult to interpret. Intuitively the content validity should be high. The items selected came from a pool of concepts covered by the course. Test performances should be relatively free from the influence of extraneous variables. Free association responding is related to verbal and ideational fluency and to vocabulary, but within the bounds imposed by restricted course content their effect should be minimal. Then too, all three are contributing factors in academic achievement and hardly qualify as extraneous variables. Students were for the most part able to provide relevant associates to the stimulus terms which also lends support to the contention of high content validity.

Correlations with subjective scores are, with the exception of English, uniformly positive, although only three are significant. The exception results in part from the extreme difficulty of the English examination and its resultant failure to discriminate, and also from the emphasis on stylistic considerations in evaluating essays on an English final examination. It seems reasonable to conclude that association tests and subjective examinations do overlap in the function or functions measured, although the degree of overlap appears to be somewhat smaller than anticipated. Several factors may have served to limit the value of the coefficients; first the heterogeneity of items grouped under the heading 'subjective', and second the fact that several finals covered material from the mid-term till the semesters end, while the word-associations covered the entire semester.

While it seems likely that word-associations and subjective examinations measure a common function (or common functions) the specific function tapped remains unclear. It is quite clear that a substantial proportion of the variance in their joint distributions is left un-

accounted for. It seems unlikely that the variance not accounted for by the correlation between word-association scores and subjective test scores is in fact error variance. This would require the paradoxical interpretation that error was reliably measured. Perhaps a factor analysis including word-associations and various types of objective and subjective test items would provide an answer to this question.

The correlations between the word-association tests and overall GPA taken as a measure of general academic achievement are best described as mixed. English (-.10), Psychology (-.15), Political Science (-.21), and Botany (-.08), all yielded negative correlations, while in the remaining five courses the coefficients were positive. Item difficulty data may account for English and Psychology but Political Science and Botany cannot be accounted for in the like manner. What makes the situation even more curious is the fact that GPA is essentially a composite score the primary component of which is subjective. If associations correlate with subjective tests, then they should also correlate with a composite based primarily on subjective measures. The curricular requirements generate a relatively homogeneous background, so that the participating students shared a common core of courses upon which the GPA was based. This homogeneity should be particularly evident for underclassmen, but all four negative coefficients were obtained in introductory courses. The issue is further confounded in that in two of these four cases the students performance on the subjective final examination was significantly correlated with GPA. Certainly additional research is needed to clarify the relationship of association test performance to GPA.

Correlation between the regular final and GPA were, with the exception of psychology, and English, high and positive. This finding is not surprising in that academic performance tends to be relatively constant. And also in view of the fact that the GPA is essentially a composite score based on previous performance on similar testing procedures.

Taken in total the validity estimates obtained are encouraging. Correlations with subjective criteria are of sufficient magnitude to suggest that associations and essays do overlap. That the overlap is not complete is not surprising in view of the stylistic considerations which typically contribute to an essay grade: correlations with GPA, though mixed, seem to indicate that associative measures are not consistently included in academic measurement. Few would contend that contemporary measuring devices tap all or even the greater part of relevant cognitive functioning. That associations are not highly

correlated with GPA may be a strength of the technique rather than a weakness.

Conclusions:

Certainly the primary conclusion to be drawn from this research is that much more research is needed to establish the utility of the word-association technique as a measure of academic achievement. The reliability of the procedure for non-quantitative undergraduate courses seems assured, providing item difficulty is moderate (a requirement for any effective test). The coefficients obtained range from .53 which is attributable to item difficulty, to .89, which is approaching the better standardized tests. The mean reliability even with the item difficulty extremes included is .75. While no normative data on the reliability of typical short teacher-made tests is available, the emphasis on strenghtening such tests in contemporary measurement texts testifies to their lack of psychometric rigor. The word-association technique provides a highly reliable, easily constructed alternative to other testing procedures.

Validity coefficients suffer somewhat from inadequate control of criterion measures. The investigator was unable to dictate either the form or the content of the regular final examinations used as subjective criterion against which the word-association test was correlated. The variety of testing procedures classified under the general rubric subjective, and the fact that some of the finals covered less than their associative counterparts undoubtedly served to diminish some of the coefficients. The small numbers enrolled in the courses made very high coefficients necessary to attain statistical significance. Despite the shortcomings, seven of the nine coefficients are in the predicted direction (three significant at the .05 level or beyond). Furthermore, the two coefficients which deviate from expectations are associated with the extremes of item difficulty. The word-association test is related to the essay examination, but further research which divorces content from style is needed to determine the degree of the relationship.

In general it seems fair to conclude that the word-association technique is of sufficient reliability and validity to warrant further investigation. At worst it provides a useful adjunt to established testing procedures, and one which seems to tap a largely untapped function. At best it may provide an "objective essay", a way to measure essay content without the normal confounding with style. While it may be condemning it with faint praise, the association test is certainly no worse

than other contemporary testing procedures.* Further research may indicate that it is significantly better.

---

* in terms of its psychometric characteristics.

BIBLIOGRAPHY

Bloom, B.S. (Ed.). Taxonomy of Educational Objectives. New York: David McKay Co. 1956.

Bousfield, W.A., The occurrence of clustering in the recall of randomly arranged associates. J.Gen. Psychol., 1953, 49, 229-240.

Cartter, A. Future faculty: needs and resources. In Lee (Ed.), Improving College Teaching. Washington, D.C.: American Council on Education, 1967, 113-135.

Cofer, C.N. Conditions for the use of verbal associations. Psychol. Bull., 1967, 68 1-12.

Cramer, Phebe, Word Association. New York: Academic Press, 1968.

Creelman, Marjorie B. The Experimental Investigation of Meaning, New York: Springer Pub. Co. 1966.

Deese, J. The Structure of Associations in Language and Thought. Baltimore: The Johns Hopkins Press, 1965.

Freedman, J.L., and Mednick, S.A. Ease of attainments of concepts as a function of responses dominance variance. J. exp. Psychol., 1958, 55, 463-466.

Mager, R.F. Preparing Instructional Objectives. Palo Alto, Calif, 1962.

Milton, E.O. The state of the establishment. In Milton and Shoben (Eds.), Learning and the Professors. Athens, Ohio: Ohio University Press, 1968, 1-9.

Noll, V.H., Introduction to Educational Measurement, 2nd Ed. Boston: Houghton Mifflin Co. 1965.

Payne, D.A. The Specification and Measurement of Learning Outcomes. Waltham, Mass: Blaisdell Pub. Co., 1968.

Remmers, H.H., Gage, N.R., and Rommel, J.F. Measurement and Evaluation. New York: Harper and Row, 1965.

Underwood, B.J., and Richardson, J. Verbal concept learning as a function of instructions and dominance level, J. exp. Psychol., 1956, 51, 229-238.

Verplanck, W.W., and Krueger, R.N. Associations, connected discourse, and "knowledge." Paper read at Psychonomic Society Convention, Chicago, October, 1965.

Verplanck, W.W. Personal communication, 1968.

Verplanck, W.W., Caplan, J.R., Jennings, W.N., and Krueger, R.N. The relationship between the content of short essays on a subject and associations to that subject: more on "knowledge." Paper read at Psychonomic Society Convention, St. Louis, October, 1966.

Verplanck, W.S. Unaware of where's awareness. In Erickson (Ed.), Behavior and Awareness. Durham, N.C.: Duke University Press, 1962, 130-158.

Woodworth, R.S., and Schlosberg, H. Experimental Psychology, New York: Henry Holt and Co, 1960.

APPENDIXES

Appendix A
Instructor Orientation Material

Association Test Instructions

In the space alotted on these pages, briefly present those associations which you make to each item which are most relevant to the subject matter of (course and number). A pinpointing word or phrase is all that is necessary to demonstrate that you know what you are writing about, and that you could write much more, if given time. Do Not repeat a word or term in response to a single stimulus word. Do Not write a sentence. Do Not guess!

Association test format

Binet	Columbus	Laissez Faire
Binet	Columbus	Laissez Faire
Binet	Columbus	Laissez Faire
Binet	Columbus	Laissez Faire

Association Test Scoring

Stimulus word: Binet

responses:		
Intelligence Test	+2	A good association, clearly demonstrates grasp of material
French psychologist	+1	Acceptable, general, reasonable guess
Stephen Vincent	0	Blank, irrelevant, out of context
Freud's Student	-1	Positively incorrect

Appendix B
Lists of Stimulus Items for Each Course
with Item Difficulty and Standard Deviation
Table 4

Introductory Psychology 212

	Word	Mean	SD.
1.	neurosis	5.27	2.56
2.	rods & cones	6.34	2.32
3.	REM	7.48	1.10
4.	adjustment	6.62	1.57
5.	Jung	6.86	1.56
6.	libido	6.65	1.45
7.	somatotype	6.58	2.25
8.	LSD	5.72	1.77
9.	self theory	5.93	1.75
10.	necker cube	5.93	1.65
11.	aggression	6.20	1.12
12.	Psychoanalysis	6.06	1.79
13.	symptom substitution	5.34	2.47
14.	Rorschach	5.89	2.29
15.	tranquilizers	4.00	2.71
16.	conflict	4.34	2.32
17.	Ames room	6.27	1.73
18.	fovea	4.75	3.06
19.	psychosis	5.65	2.60
20.	perception	6.65	1.30

Political Philosophy 212

	Word	Mean	SD.
1.	Dialectical Materialism	5.00	1.18
2.	Fascism	5.29	2.35
3.	Vanguard of the Proletariat	6.11	1.73
4.	Constitutionalism	4.76	1.92
5.	Classical liberalism	6.11	2.74
6.	Leviathan	5.29	2.62
7.	General Will	4.76	2.93
8.	Investiture Controversy	2.76	2.59
9.	Dante's De Monarchia	5.17	2.81
10.	Stalin's contribution to Communism	3.47	2.78
11.	The Conciliar Movement	3.58	2.74
12.	Machiavelli	6.11	2.00
13.	Pre-Plato Political Thought	3.52	2.57
14.	Levellers	3.47	3.08
15.	Utilitarianism	6.29	2.12
16.	John Stuart Mill	5.64	1.97
17.	Plato's the Laws	5.35	2.67
18.	John Locke	6.70	1.37
19.	Stoicism	4.88	3.08
20.	Bodin's concept of Sovereignity	2.05	2.05

Appendix B (contd.)

History 212

	Word	Mean	SD.
1.	deal	6.35	1.42
2.	radicals	5.00	2.03
3.	Kingfish	4.35	2.92
4.	normalcy	4.00	2.60
5.	Sinclair	6.23	1.40
6.	embargo	6.17	1.86
7.	Homer	2.11	3.08
8.	Maine	4.41	2.40
9.	warzone	5.76	2.46
10.	resumption	3.05	3.45
11.	Tweed	3.94	1.56
12.	Johnson	6.17	1.72
13.	silver	3.70	1.77
14.	Grant	3.23	1.82
15.	Wormley House	0.70	3.90
16.	populism	4.82	1.73
17.	trust	4.47	1.50
18.	Panama	4.41	1.66
19.	round robin	1.47	4.25
20.	Calvin	4.94	1.60

English 112

	Word	Mean	SD.
1.	Greenwich Observatory	2.64	2.29
2.	Sphinx	1.70	2.17
3.	Petrachan	2.88	2.94
4.	Gloucester	4.29	3.08
5.	A blind king	2.52	3.70
6.	preexistence	0.00	1.12
7.	The perfect detonator	2.05	2.08
8.	toothpaste	4.23	3.44
9.	metaphysical	0.76	1.92
10.	onion cellar	2.47	2.45
11.	Court of Justice	0.41	1.24
12.	a double	1.82	2.84
13.	pillbox	1.41	2.42
14.	Confederate cavalry	0.94	1.68
15.	lass of Augrim	0.23	0.75
16.	telescope	2.94	2.64
17.	national guard	0.35	1.54
18.	carving knife	1.82	2.22
19.	an illegitimate son	2.82	1.70
20.	four skirts	4.29	2.26

Appendix B (contd.)

Philosophy 252

	Word	Mean	SD.
1.	Voluntarism	6.33	1.68
2.	Categorical Imperative	5.00	2.28
3.	interest theories	2.50	2.78
4.	Ring of Cyges	4.05	3.00
5.	Stoicism	2.66	2.40
6.	Autonomy of the Will	1.38	2.09
7.	Cyrenaicism	4.05	3.64
8.	Ethical Intuitionism	3.00	3.20
9.	Slave-morality	3.38	3.07
10.	Principle of Universalizability	1.94	2.82
11.	"Is-Ought" problem	2.83	1.89
12.	Principle of Utility	6.16	1.89
13.	Philosophic Wisdom	0.88	0.91
14.	Hedonistic Calculus	4.61	3.01
15.	Problem of Evil	1.38	1.94
16.	Platonic Forms	2.94	1.86
17.	Emotivism	3.66	3.44
18.	Casuistry	3.72	1.65
19.	Ethical Rationalism	4.44	2.97
20.	The Sanctions of Utility	1.11	2.09

Economics 212

	Word	Mean	SD.
1.	social imbalance	5.00	1.88
2.	"workable" competition	3.00	2.28
3.	regulated monopoly	3.77	2.34
4.	excess supply	3.55	1.54
5.	elasticity of demand	2.72	2.68
6.	dollar devaluation	3.33	1.68
7.	mutual interdependence	4.50	2.36
8.	bilateral monopoly	3.61	2.64
9.	economic profit	2.50	2.67
10.	comparative advantage	3.27	2.52
11.	GATT	2.55	2.25
12.	dollar glut	2.38	2.09
13.	pure competition	6.77	1.86
14.	parity	3.72	1.33
15.	long run ATC curve	1.38	2.38
16.	economic rent	2.33	2.38
17.	general equilibrium	1.22	1.11
18.	demand curve	3.38	1.81
19.	Taft-Hartley	3.61	2.20
20.	marginal revenue product	3.16	2.29

Appendix B (contd.)

Education 212

	Word	Mean	SD.
1.	Social Reconstructionism	2.09	1.70
2.	Great Books Program	2.43	1.42
3.	Nausea	2.31	2.16
4.	Universals	2.43	2.06
5.	Mental discipline	3.31	1.54
6.	Scholasticism	2.06	2.03
7.	Hegel's dialectic	2.03	2.02
8.	Apperception	1.62	2.05
9.	Socratic Method	2.59	2.34
10.	Wittgenstein	3.25	2.75
11.	Congruence theory	2.62	2.10
12.	Pragmatism	4.78	2.30
13.	Faculty psychology	3.09	2.00
14.	Teleology	1.18	1.56
15.	Categorical imperative	1.75	2.00
16.	Determinism	2.12	2.13
17.	tabula rasa	1.53	2.30
18.	Allegory of the Cave	1.96	1.89
19.	Form-Matter Hypothesis	1.87	2.24
20.	a posteriori	2.06	1.93

Religion and Culture 332

	Word	Mean	SD.
1.	Faces	4.54	1.29
2.	Mundane World	4.04	1.52
3.	Story	3.62	2.28
4.	Boundaries	3.91	1.69
5.	Home	4.75	1.54
6.	Baldicer	5.16	2.44
7.	Unimagining American	4.25	1.96
8.	Mouth	5.16	2.14
9.	Orestes	4.04	2.44
10.	Going Abroad	4.62	1.84
11.	Body	5.29	1.66
12.	Bacchae	3.91	2.10
13.	Upright Posture	3.54	1.93
14.	Left-Hand Knowing	3.58	2.60
15.	Lottery	3.66	2.01
16.	Time	5.04	1.73
17.	Actor	3.08	2.13
18.	Primordial	2.58	2.30
19.	Responsibility	2.45	2.25
20.	Ritual	3.45	2.09

Appendix B (contd.)

Botany 102

	Word	Mean	SD.
1.	Stomates	6.00	2.00
2.	Sporophyte	5.73	2.20
3.	Endodermis	5.62	2.39
4.	Plant distribution	5.64	2.48
5.	Carpel	5.57	2.57
6.	Archegonium	5.00	2.91
7.	Berry	4.12	2.90
8.	Cambium	3.91	2.68
9.	Auxin	4.23	2.22
10.	primary growth	5.07	2.27
11.	Seed	5.75	2.41
12.	Bryophyte	5.37	2.92
13.	Annulus	4.01	3.33
14.	Limiting factor	3.41	3.18
15.	Osmosis	4.37	2.86
16.	Fungi	4.92	2.70
17.	OOgamy	5.23	2.66
18.	Respiration	4.82	3.04
19.	Xylem	5.23	2.91
20.	Gameophyte	5.26	2.84

Appendix C
Sample Test

ASSOCIATION TEST INSTRUCTIONS

In the space alotted on these pages, briefly present those associations which you make to each stimulus word which are most relevant to the subject matter of History 212. A pinpointing word or phrase is all that is needed to demonstrate that you know what you are writing about, and that you could write much more, if given time. Do not repeat a response, do not write sentences, and do not guess.

Example

Lusitania WWI
Lusitania submarines
Lusitania Wilson
Lusitania strict accountability

Scoring

Your responses will be scored as follows:

+2 A good association, clearly indicates grasp of the material.
+1 Acceptable, but not very informative, Overly general, a good guess
0 A blank, out of context, irrelevant
-1 Clearly incorrect

Appendix C (contd.)

History 212 Exam Name_____

deal	radicals	Kingfish
deal	radicals	Kingfish
deal	radicals	Kingfish
deal	radicals	Kingfish
normalcy	Sinclair	embargo
normalcy	Sinclair	embargo
normalcy	Sinclair	embargo
normalcy	Sinclair	embargo
Homer	Maine	warzone
Homer	Maine	warzone
Homer	Maine	warzone
Homer	Maine	warzone
	resumption	
	resumption	
	resumption	
	resumption	

Appendix C (contd.)

History 212 Exam Name_____

Tweed	Johnson	silver
Tweed	Johnson	silver
Tweed	Johnson	silver
Tweed	Johnson	silver
Grant	Wormley House	populism
Grant	Wormley House	populism
Grant	Wormley House	populism
Grant	Wormley House	populism
trust	Panama	round robin
trust	Panama	round robin
trust	Panama	round robin
trust	Panama	round robin
	Calvin	
	Calvin	
	Calvin	
	Calvin	

Appendix D

Word Count

Source: Political Phil. 432
Date: 1971
Stimulus Word: Dante's De Monarchia

Associations	Score	Positions 1	2	3	4	Total Freq.	Total Score
Nation-state	2	1				1	2
Unity	2		1			1	2
anti-Church	2			2		2	4
Federation	2				1	1	2
Nationalistic	2	2				2	4
Church-State	2	1				1	2
secular control	2		1			1	2
pro state	2	1				1	2
Italy	2		1			1	2
Religion harmful	2			1		1	2
Monarchy rule	2	1			1	2	4
Nationalism	2		1	1		2	4
Italy center of world	2			1		1	2
Church's power lessen	2				1	1	2
supportive king	2		1			1	2
world gov't	2	6				6	12
emphasis on order	1		1			1	1
secular	2			1		1	2

Appendix D (contd.)

Word Count

Source: Political Phil.432
Date: 1971
Stimulus Word: Dante's De Monarchia

Associations	Score	Positions 1	2	3	4	Total Freq.	Total Score
unified Italy	2	1				1	2
International Law	2		1			1	2
One government	2			1		1	2
Aristotle's arguement	2				1	1	2
Federal system	2		1			1	2
Italian nationalism	2		3			3	6
Unification of states	2		1			1	2
Order oriented	2			1		1	2
Freedom for state, not individualism	2				1	1	2
Secularism	2				1	1	2
Christian empire	2	1				1	2
Corruption of Church	2			1		1	2
Rome as center	2				1	1	2
federalism	2		1			1	2
Catholicism	1	1				1	1
Church in state	2		1			1	2

Appendix E

Table 5

WORD- ASSOCIATION TEST PSYCHOMETRIC CHARACTERISTICS

Test Number	Course Area	Number of Subjects	Reliability Coefficients
1	Introductory Psych. [a]	25	.724[c]
2	Introductory Psych. [a]	18	.736[c]
3	Introductory Psych. [a]	133	.844[c]
4	Introductory Psych.	15	.847[c]
5	Animal Behavior [b]	20	.914[c]
6	History and Systems [b]	28	.943[c]
7	Introductory Psych. [a]	117	
8	Personality [b]	82	.873[d]
9	Physiological Psych. [b]	74	.923[d]
10	Sensation and Perception [b]	27	.912[c]
11	Statistics and Methods	19	.789[c]
12	Personality [b]	19	.909[c]
13	History and Systems [b]	24	.904[c]
14	Learning [b]	24	.975[c]
15	Testing [b]	18	.823[c]
16	Physiological Psych. [b]	28	.889[c]
17	Social Psych. [b]	19	.854[c]
18	Introductory Economics [b]	364	.885[d]

[a] Undergraduate.

[b] Graduate.

[c] Kuder-Richardson formula - 20.

[d] Spearman-Brown formula.

Appendix E (contd.)

Table 6

WORD-ASSOCIATION TEST VALIDITY COEFFICIENTS

Test Number	G.P.A.	Mult. Choice	Short Essay	Essay	Dean's List	A.C.T. Verbal	A.C.T. Composite
1	.447		.733	.602	678		
2							
3	.356	.459					
4			.329	-.013			
5	.647						
6	.455						
7	.605	.632	.647			.314	.416
8							
9							
10							
11							
12							
13							
14							
15							
16							
17							
18	.419	.562					

Appendix F

KR-20 RELIABILITY CALCULATION

Formula Adapted by E.E. Cureton - University of Tennessee
Knoxville, Tenn.

		ITEMS					
		1	2	3	4	Σx_i	$(\Sigma x_i)^2$
Students	1	6	5	7	6	24	576
	2	5	7	4	3	19	361
	3	3	4	3	-2	8	64
	4	0	4	3	6	13	169
Σx_j		14	20	17	13	64	1170
$\Sigma(x_j^2)$		70	106	83	85	344	
$(\Sigma x_j)^2$		196	400	289	169	1054	

J = 1, Number of students, I = 1,K
K = Number of items

$$KR\text{-}20 = \frac{K}{K-1}\left[1 - \frac{N\,x_j^2 - (\Sigma x_j)}{N\,x_i^2 - (\Sigma x_i)^2}\right]$$

$$= \frac{4}{3}\left[1 - \frac{4(344) - 1054}{4(1170) - (64)^2}\right]$$

$$= \frac{4}{3}\left[1 - \frac{1376 - 1054}{4680 - 4096}\right]$$

$$= \frac{4}{3}\left(1 - \frac{322}{584}\right)$$

$$= .60$$